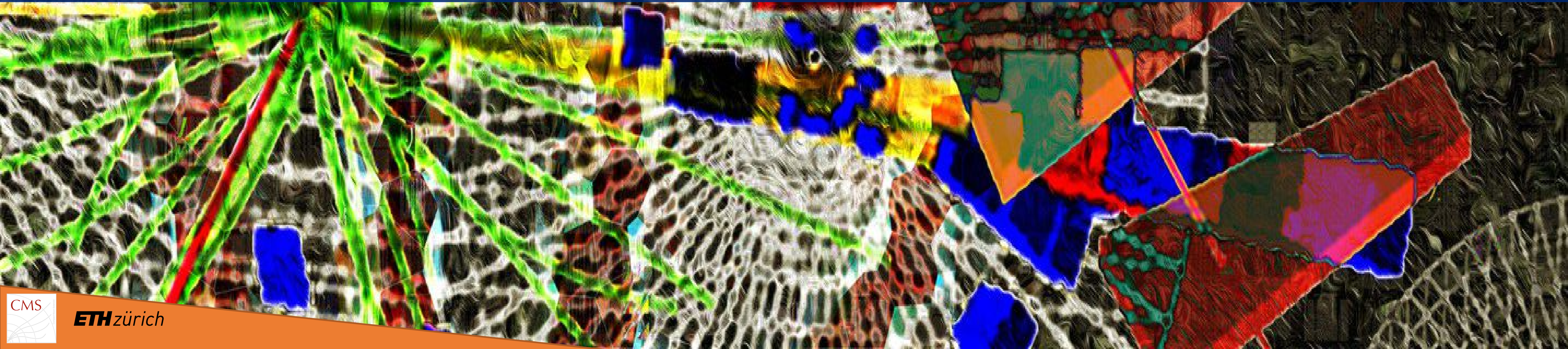
Electromagnetic Shower Corrections in CMS

X. Chang, G. Dissertori, M. Donegá, M. Galli, P. Musella, S. Pigazzini, T. Reitenspiess

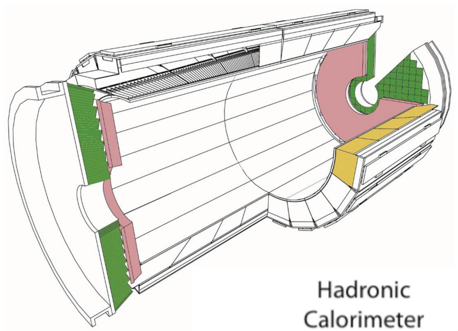
22 March 2023

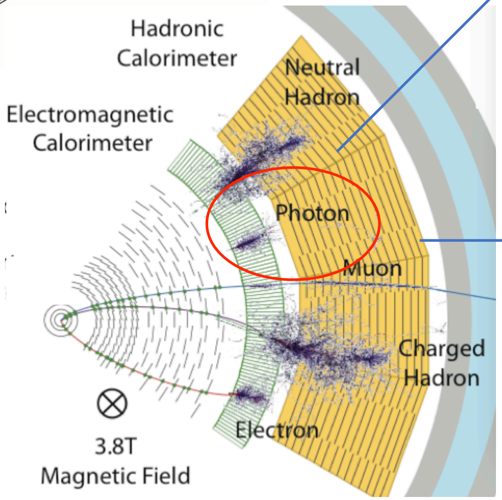


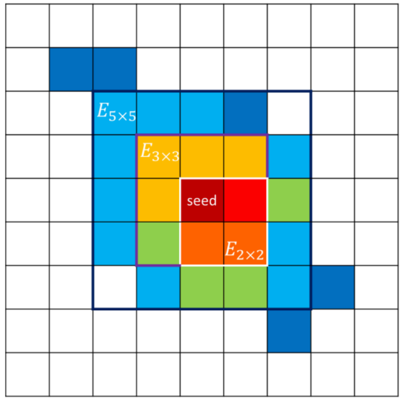
CMS ETH zürich

Simulation of Electromagnetic Variables

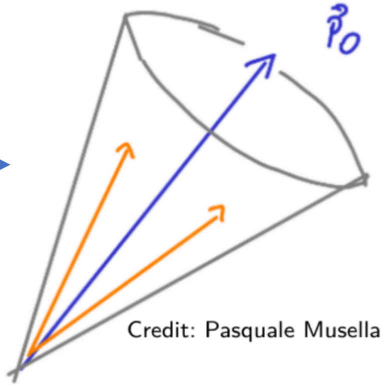
- Monte Carlo (MC) used in all the analyses with CMS data
- When it comes to analyses that use **photons** (e.g. $H \rightarrow \gamma\gamma$), the description of the **electromagnetic shower** in the ECAL is crucial:







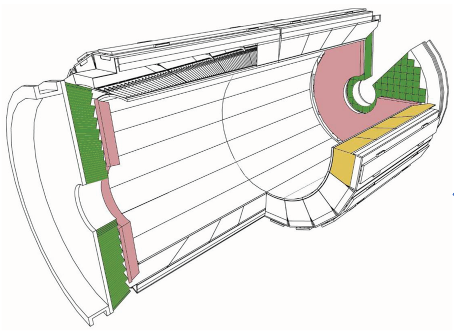
Shower shape variables: describe the shape of the EM shower cluster in the calorimeter



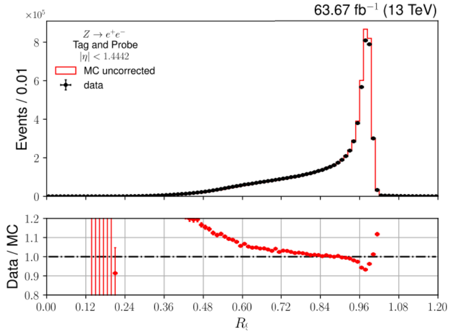
Isolation variables: characterize the activity around the object of interest

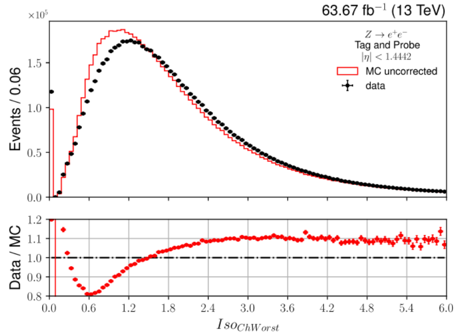
Simulation of Electromagnetic Variables

- Monte Carlo (MC) used in all the analyses with CMS data
- When it comes to analyses that use **photons** (e.g. $H \rightarrow \gamma\gamma$), the description of the **electromagnetic shower** in the ECAL is crucial:



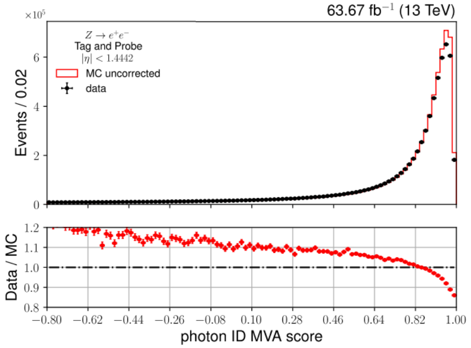
Detector aging makes it difficult to correctly simulate the shower development





Data - MC mismatch in shower shapes and isolation variables

ID MVA

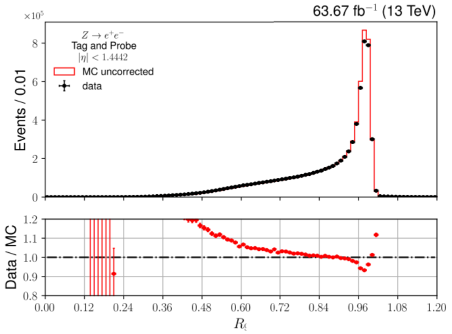


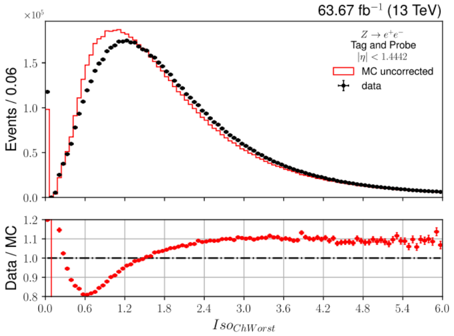
Disagreement propagated to **photon identification**…

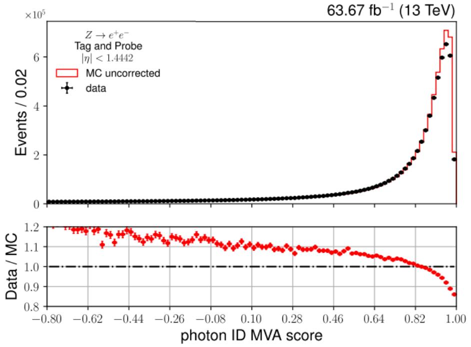
… which ultimately results in **higher systematic uncertainties**

Simulation of Electromagnetic Variables

- Monte Carlo (MC) used in all the analyses with CMS data
- When it comes to analyses that use **photons** (e.g. $H \rightarrow \gamma\gamma$), the description of the **electromagnetic shower** in the ECAL is crucial:



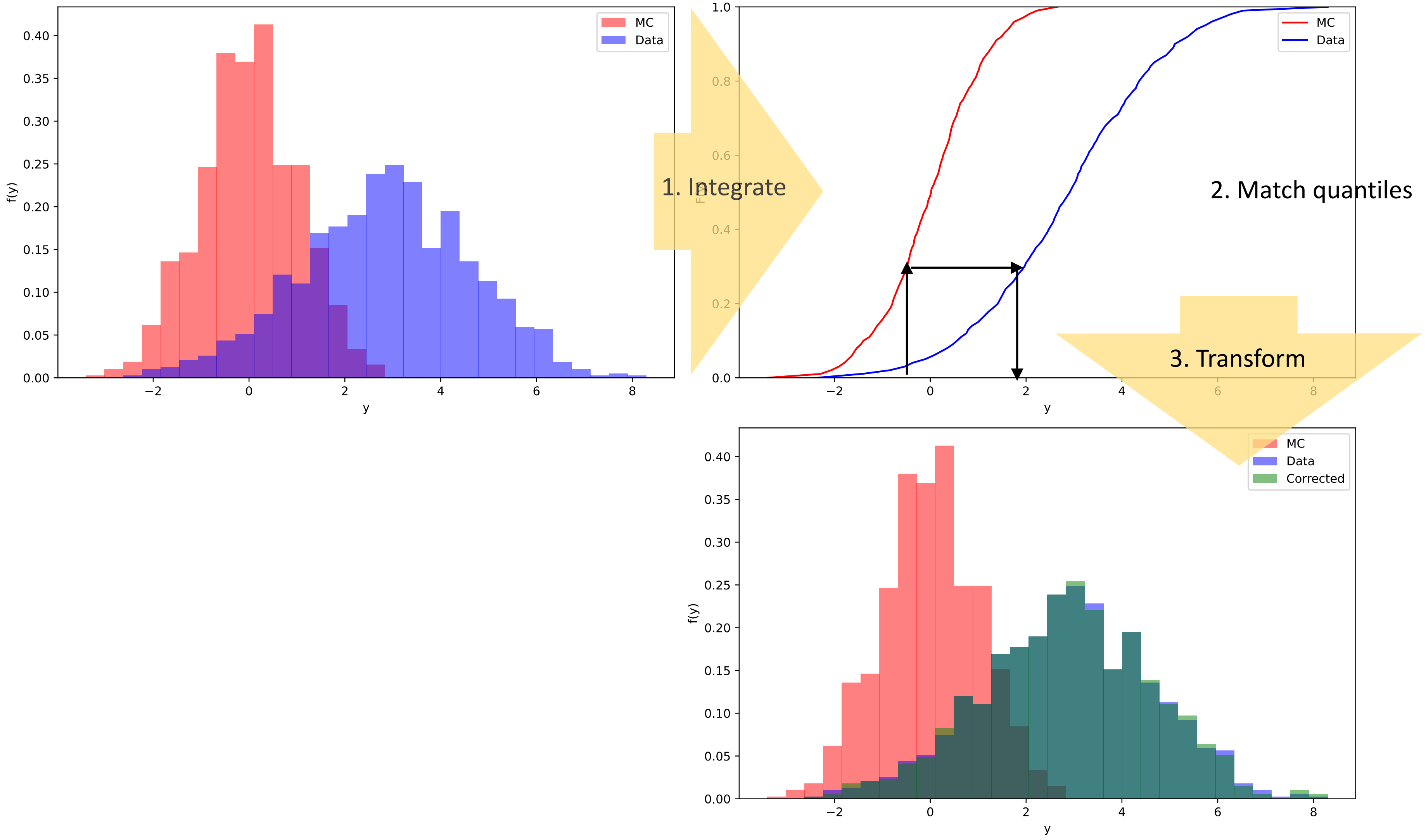




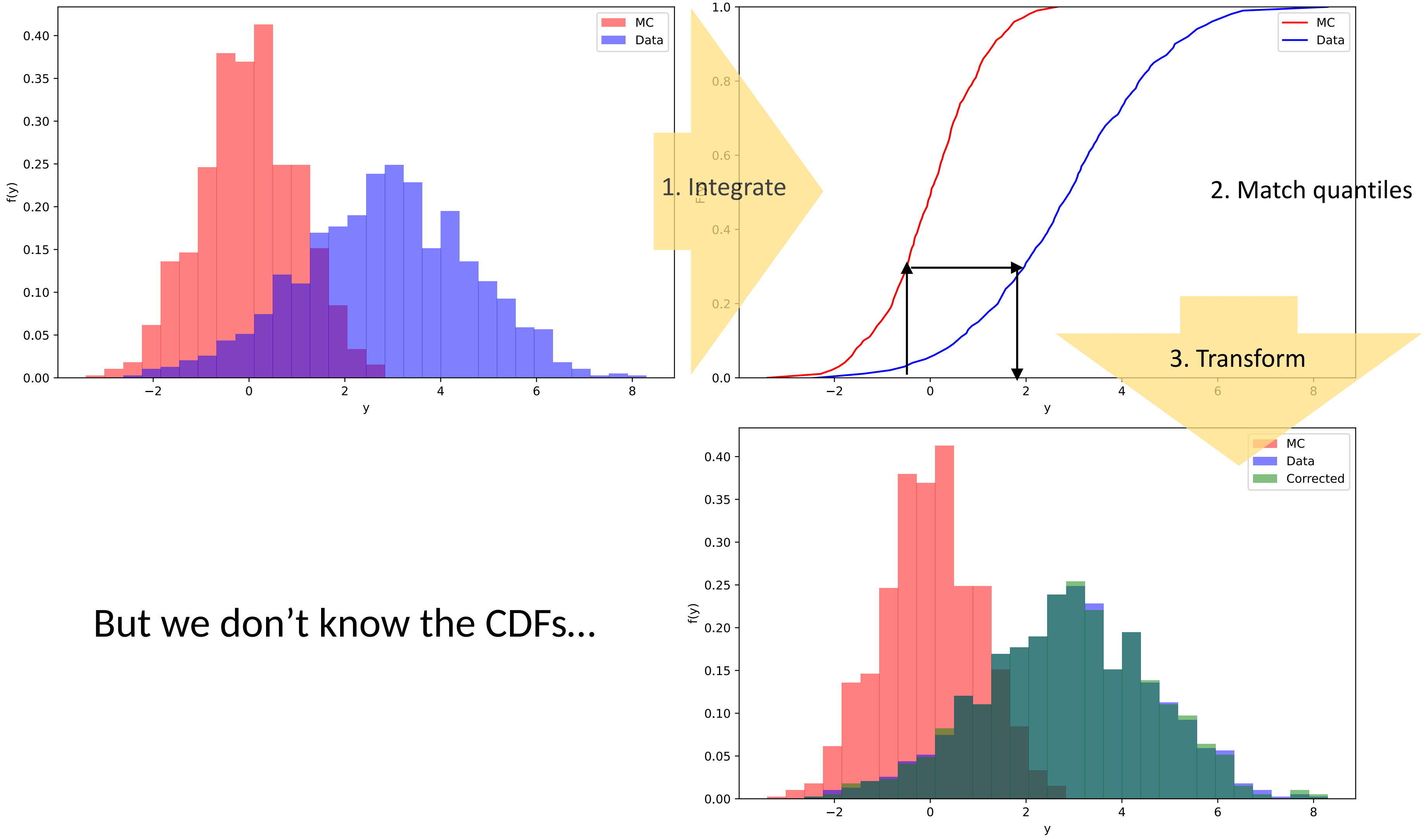
Developed a procedure called Chained Quantile Regression (**CQR**) to match MC with data (and hence decrease systematic uncertainties)

CMS ETH zürich

Quantile Morphing



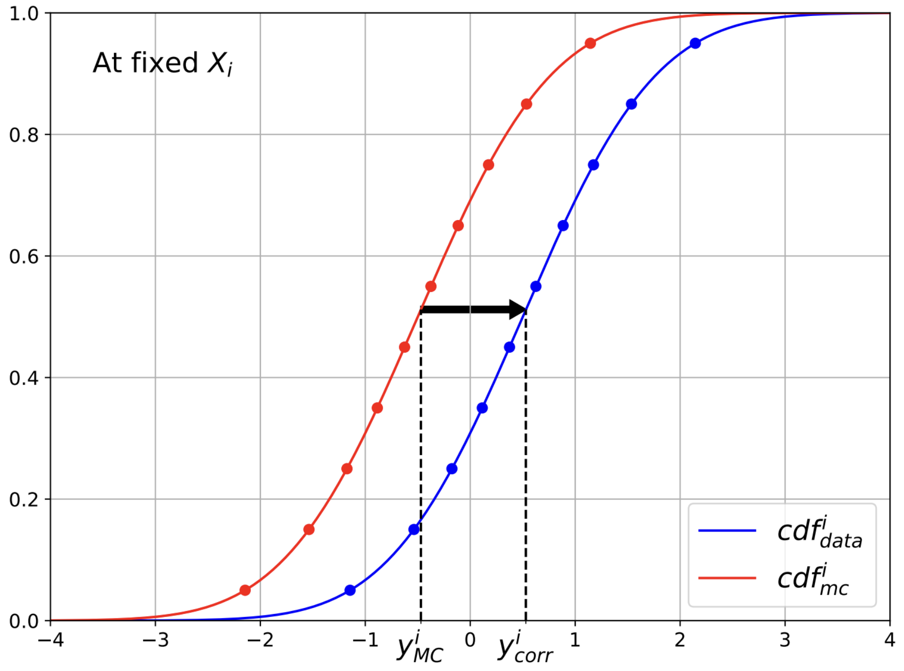
Quantile Morphing



But we don’t know the CDFs…

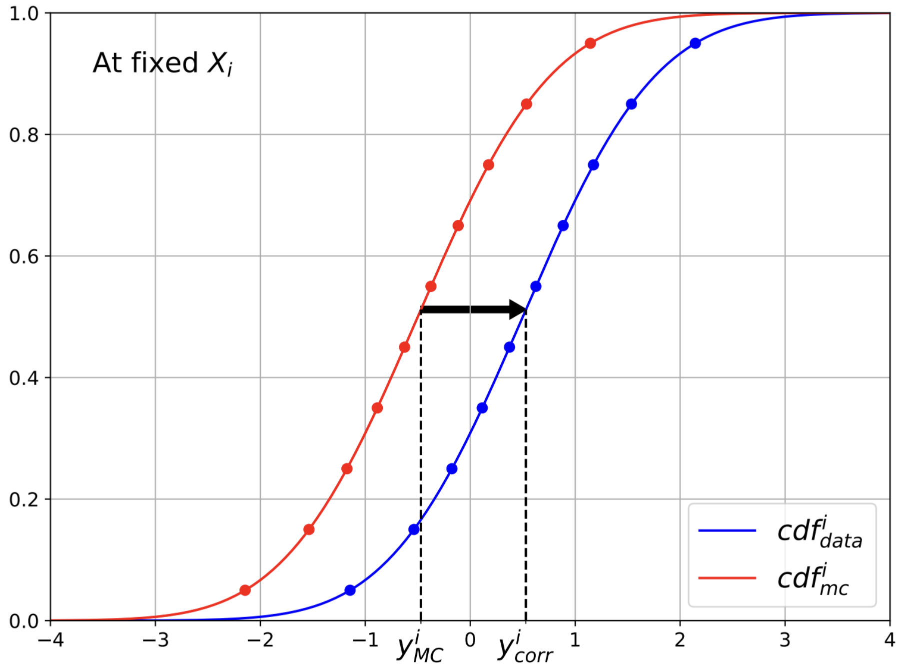
Quantile Regression

- Cumulative Distribution Function (**CDF**) of both data and MC depend on kinematic quantities $X = [p_t, \eta, \phi, \rho]$ - which describe the physics of the shower
- **Train regressors** to predict the conditional shape of CDFs using **21 quantiles**
- To correct a certain variable y_i^{MC}:
  - Find two quantiles around y_i^{MC} for data and MC
  - Use linear interpolation between the two points to obtain $cdf^{data}(y_i\,|\,X_i)$ and $cdf^{MC}(y_i\,|\,X_i)$
  - Compute $y_i^{MC,corr}$ by solving $y_i^{MC,corr} = cdf^{-1}_{data}(cdf^{MC}(y_i^{MC}\,|\,X_i))$



Quantile Regression

- Cumulative Distribution Function (**CDF**) of both data and MC depend on kinematic quantities $X = [p_t, \eta, \phi, \rho]$ - which describe the physics of the shower
- **Train regressors** to predict the conditional shape of CDFs using **21 quantiles**
- To correct a certain variable y_i^{MC}:
  - Find two quantiles around y_i^{MC} for data and MC
  - Use linear interpolation between the two points to obtain $cdf^{data}(y_i | X_i)$ and $cdf^{MC}(y_i | X_i)$
  - Compute $y_i^{MC,corr}$ by solving $y_i^{MC,corr} = cdf_{data}^{-1}(cdf^{MC}(y_i^{MC} | X_i))$



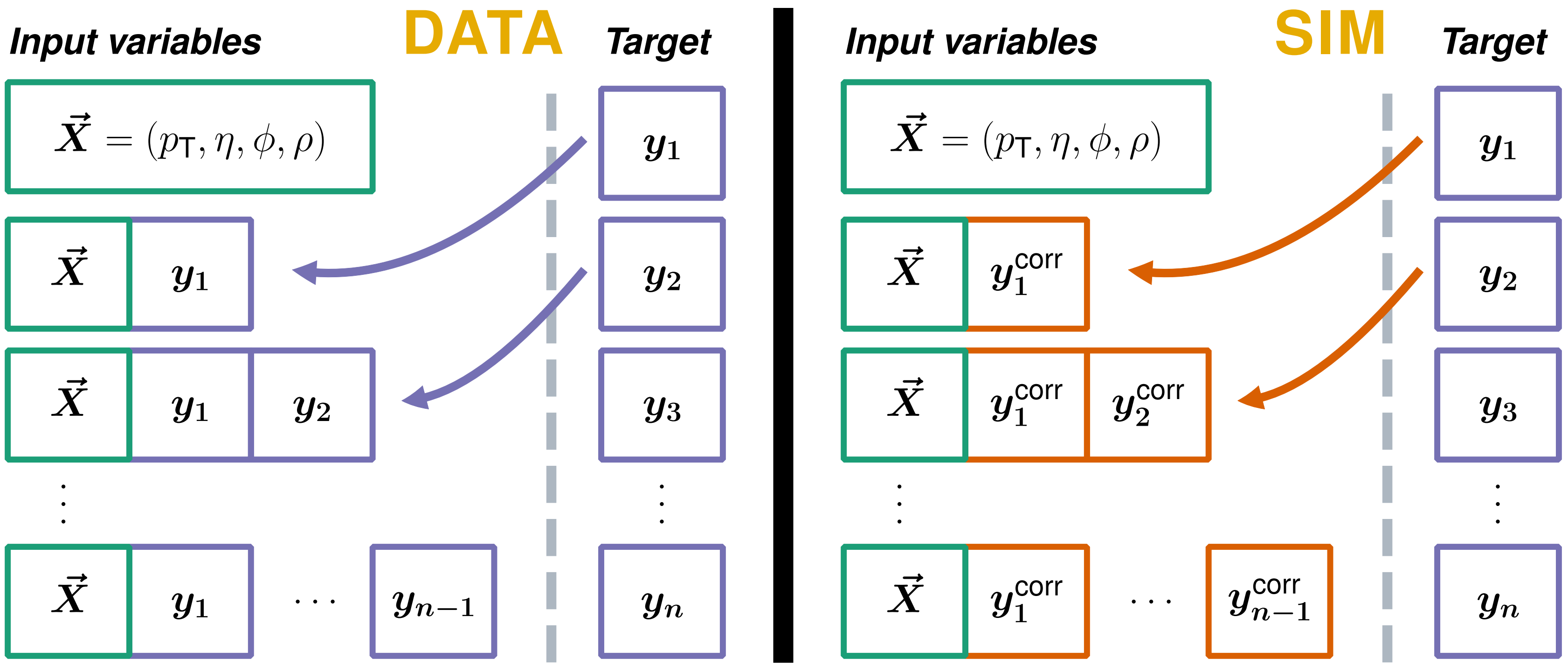
But this is not enough because the variables are correlated...

CMS

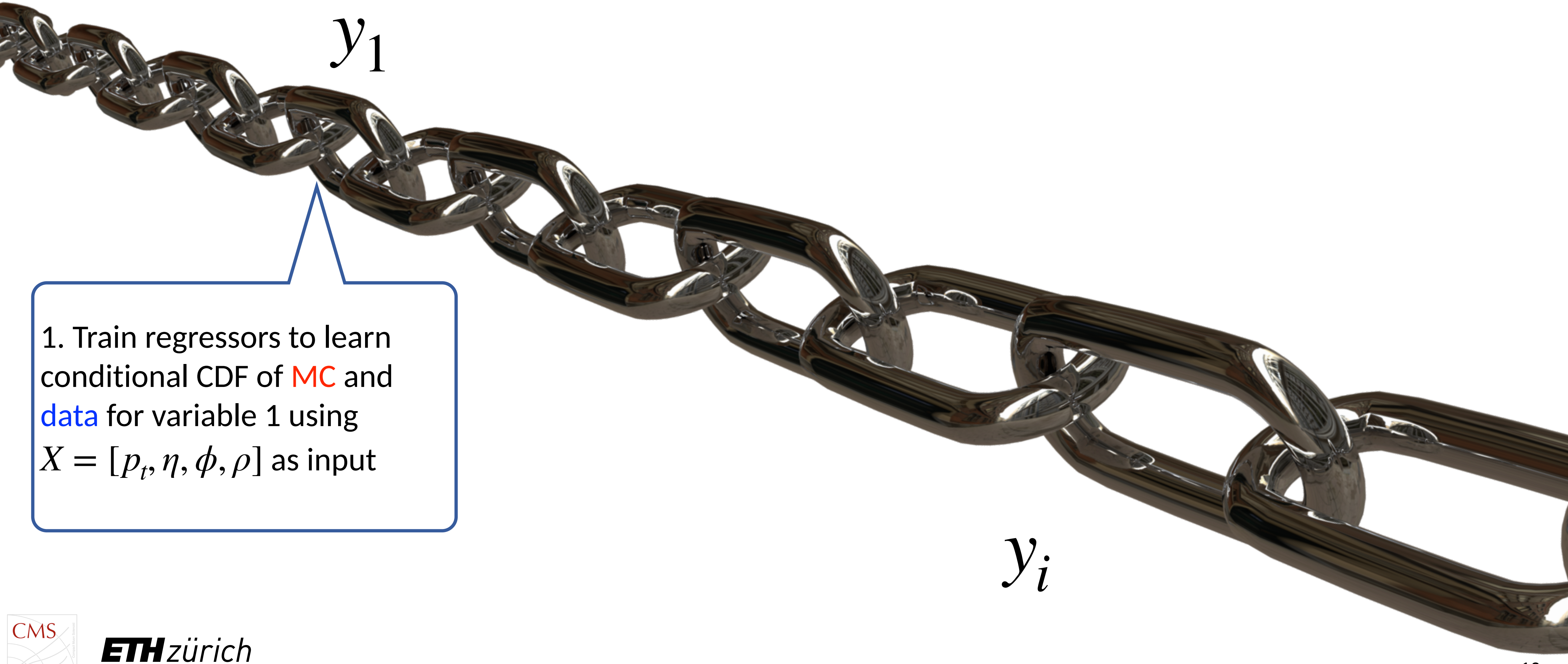
ETH zürich

Chained Quantile Regression

- In order to **catch correlations** between the variables we are correcting we need to chain them:
  - Data: for target variable y_i input variables are $X = [p_t, \eta, \phi, \rho, y_1, \ldots, y_{i-1}]$
  - MC: for target variable y_i input variables are $X = [p_t, \eta, \phi, \rho, y_1^{corr}, \ldots, y_{i-1}^{corr}]$

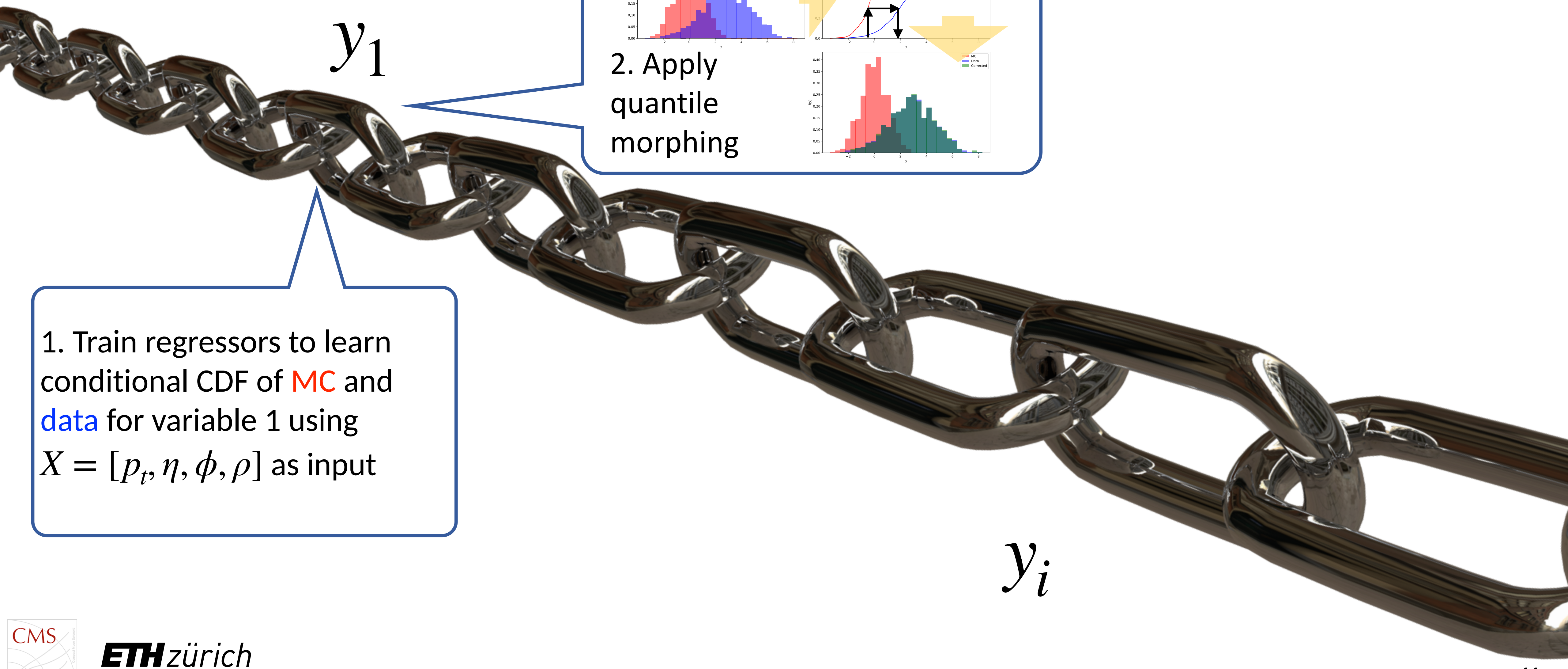


Chained Quantile Regression



1. Train regressors to learn conditional CDF of MC and data for variable 1 using $X = [p_t, \eta, \phi, \rho]$ as input

Chained Quantile Regression



y_1

2. Apply quantile morphing

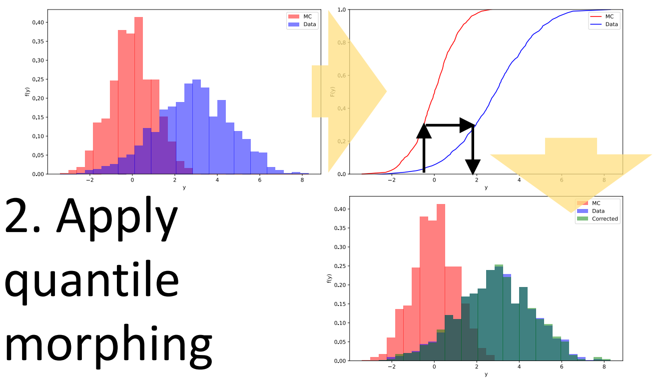
1. Train regressors to learn conditional CDF of MC and data for variable 1 using $X = [p_t, \eta, \phi, \rho]$ as input

y_i

Chained Quantile Regression

y_1

2. Apply quantile morphing



3. Repeat the procedure for variable i using

$X = [p_t, \eta, \phi, \rho, y_1, \ldots, y_{i-1}]$

(data)

$X = [p_t, \eta, \phi, \rho, y_1^{corr}, \ldots, y_{i-1}^{corr}]$

(MC) as input

1. Train regressors to learn conditional CDF of MC and data for variable 1 using $X = [p_t, \eta, \phi, \rho]$ as input

y_i

What are these regressors?

In its first implementation (still used in most $H \rightarrow \gamma\gamma$ analyses) **one BDT per quantile** was trained:

21 BDTs

x 9 variables

x 2 samples

x 2 detector parts

+ (...)

= many BDTs!

Computationally expensive and time consuming!

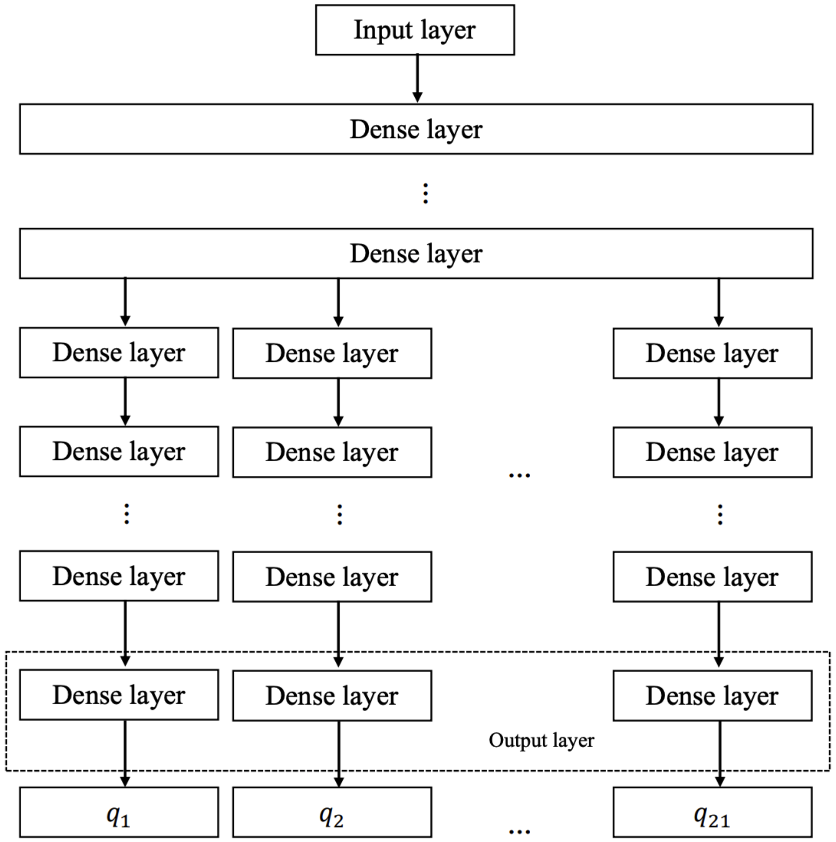
What are these regressors?

In its first implementation (still used in most $H \rightarrow \gamma\gamma$ analyses) **one BDT per quantile** was trained:

21 BDTs
x 9 variables
x 2 samples
x 2 detector parts
+ (...)
= many BDTs!

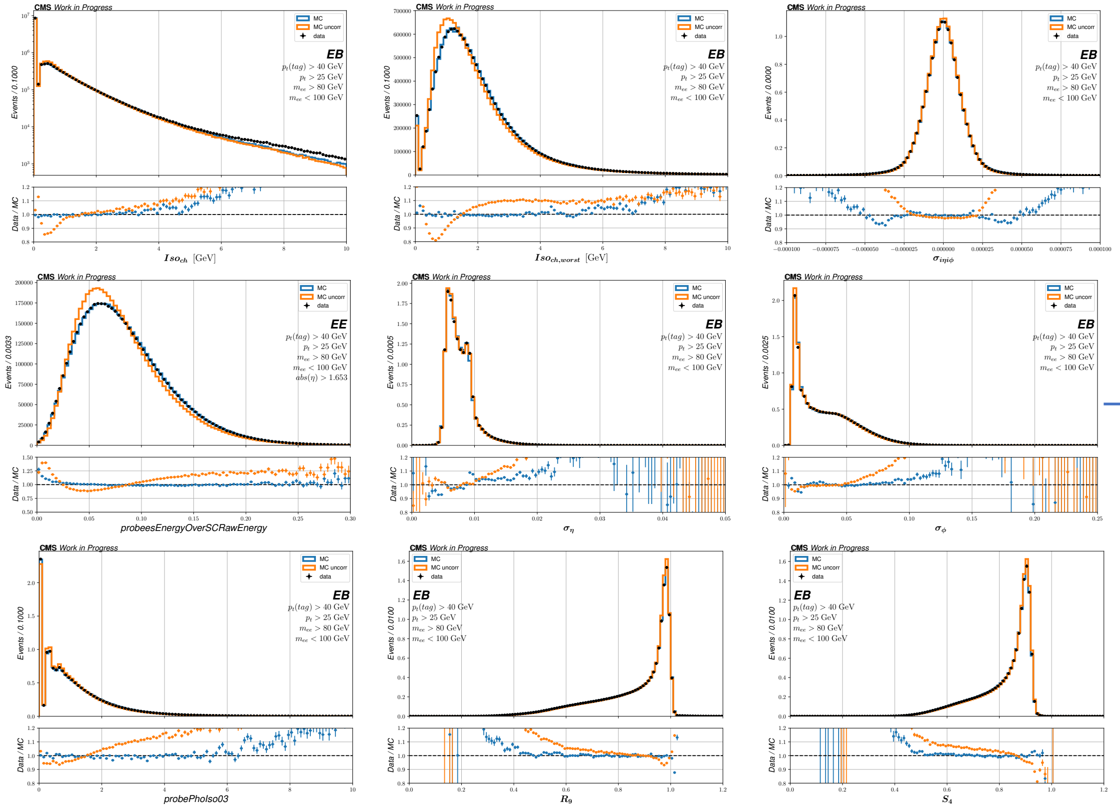
Computationally expensive and time consuming!

In a second iteration of the work the 21 BDTs of each variable were replaced by a single **quantile regression neural network**:

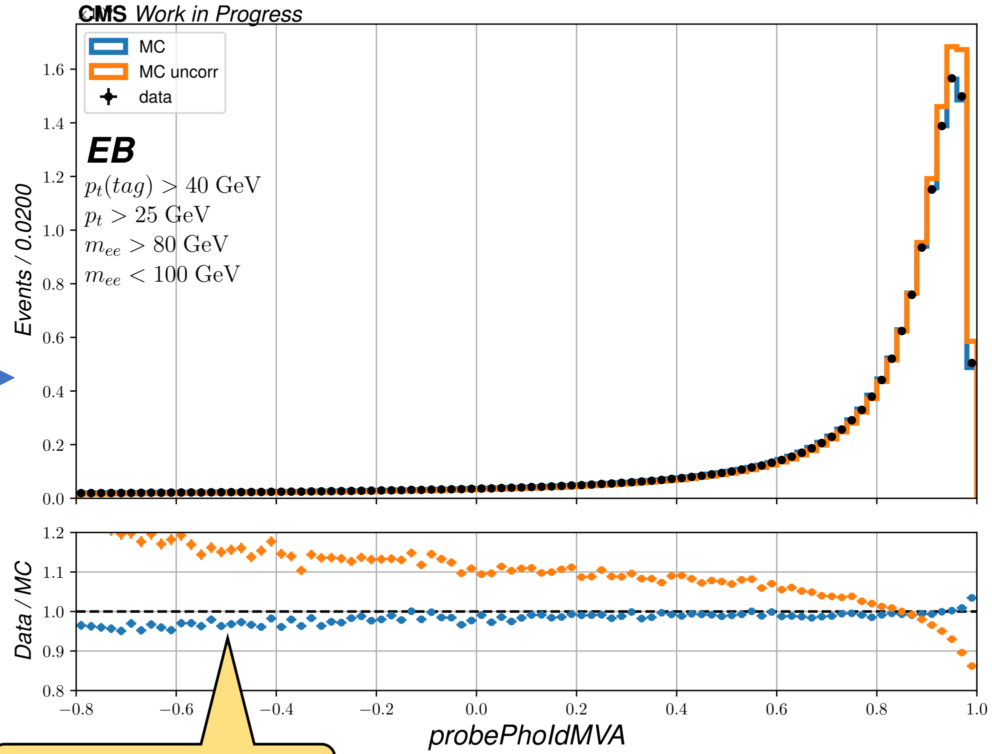


CMS ETH zürich

Corrected Distributions



ID MVA

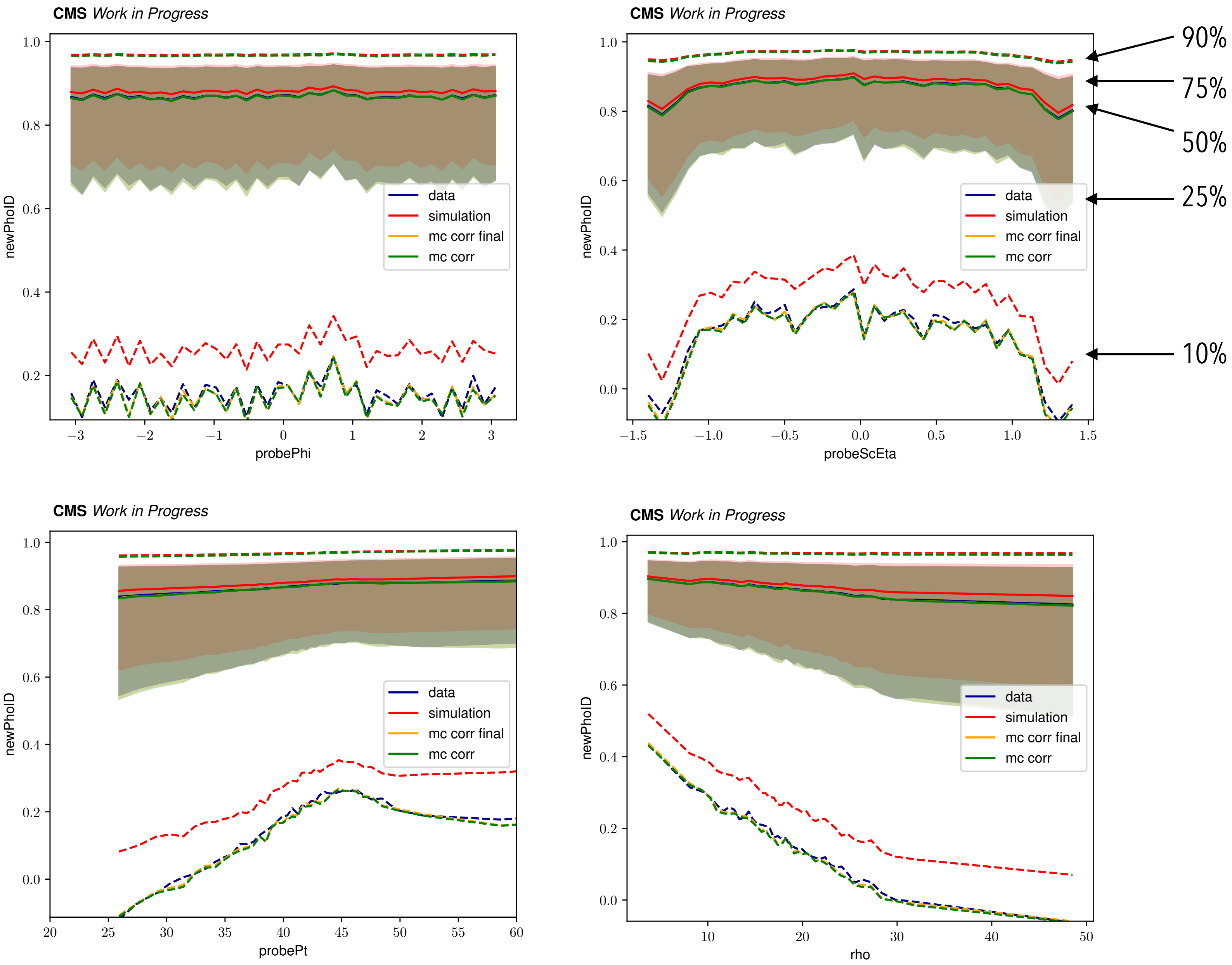


Ratio closer to 1 = **better agreement**

CMS

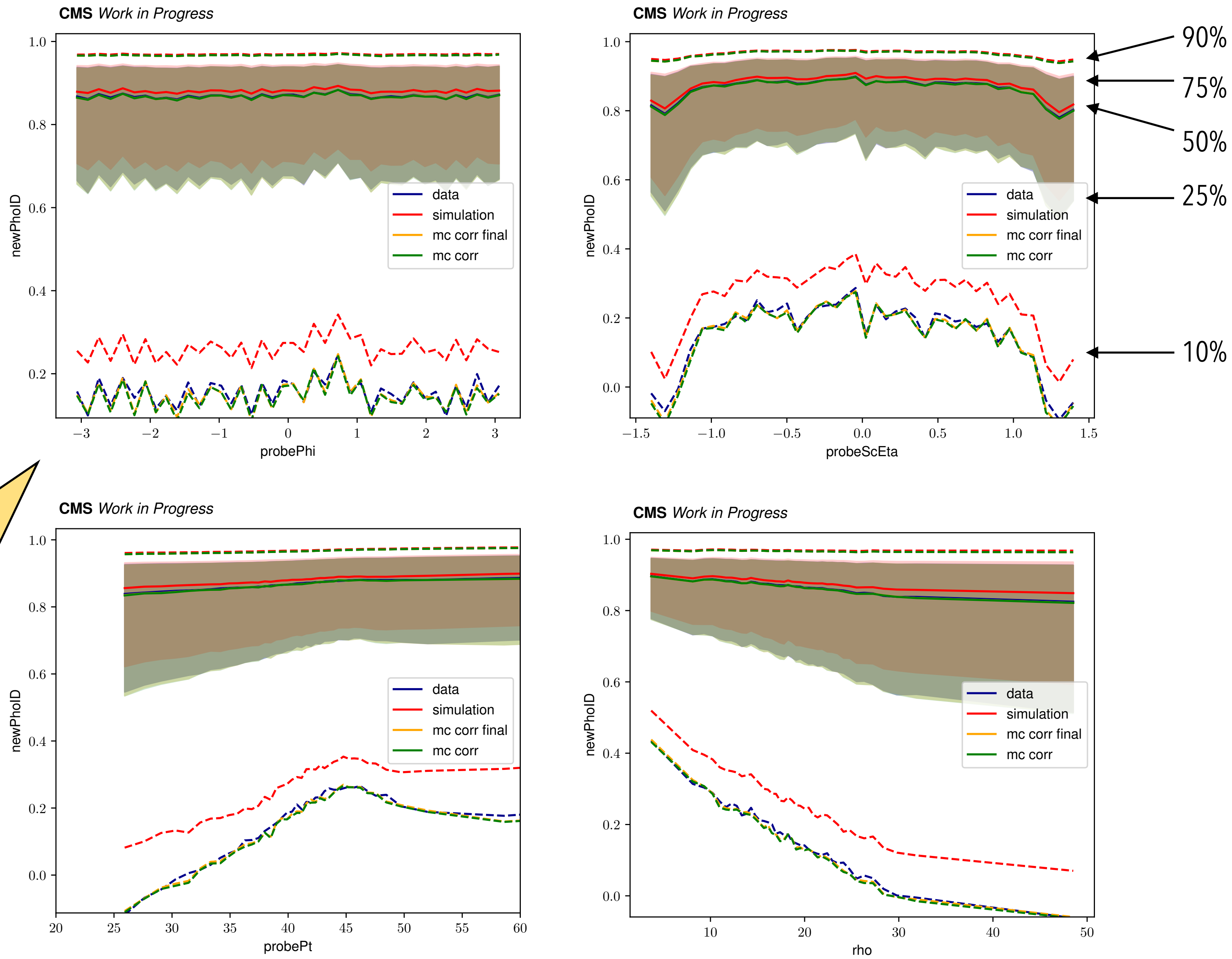
ETH zürich

Profiles



CMS Work in Progress
newPhoID
probePhi
data
simulation
mc corr final
mc corr
CMS Work in Progress
newPhoID
probeScEta
data
simulation
mc corr final
mc corr
90%
75%
50%
25%
10%
CMS Work in Progress
newPhoID
probePt
data
simulation
mc corr final
mc corr
CMS Work in Progress
newPhoID
rho
data
simulation
mc corr final
mc corr

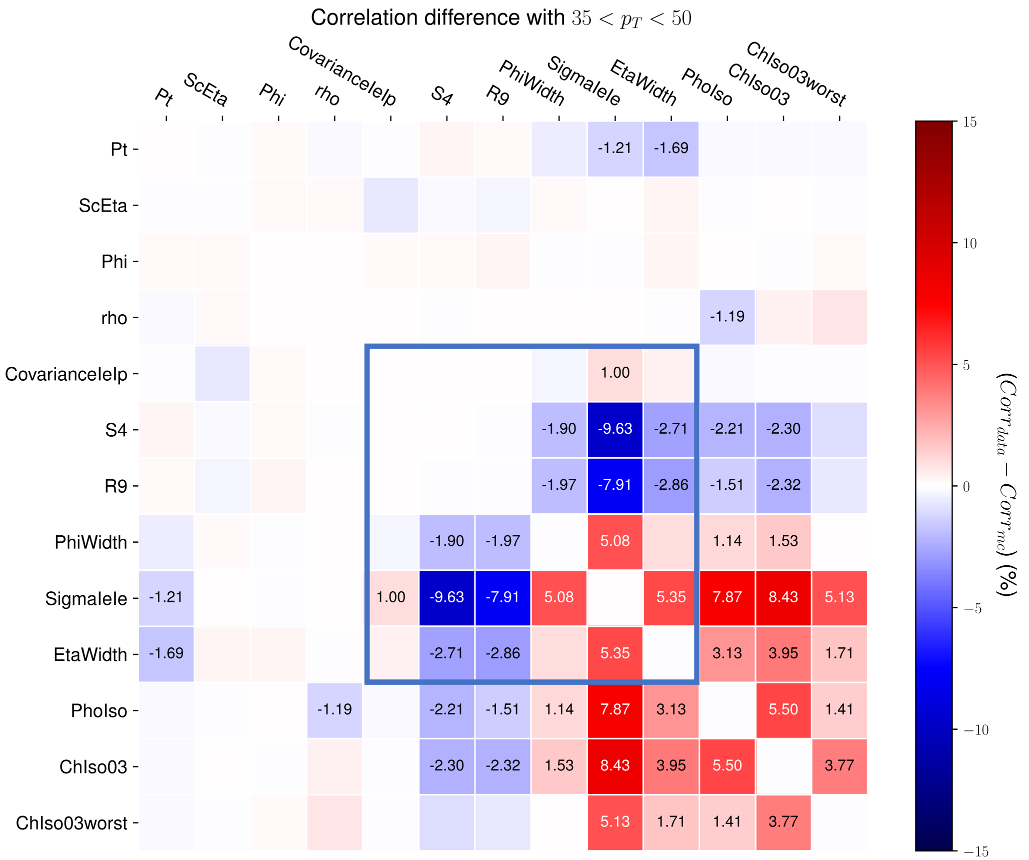
Profiles

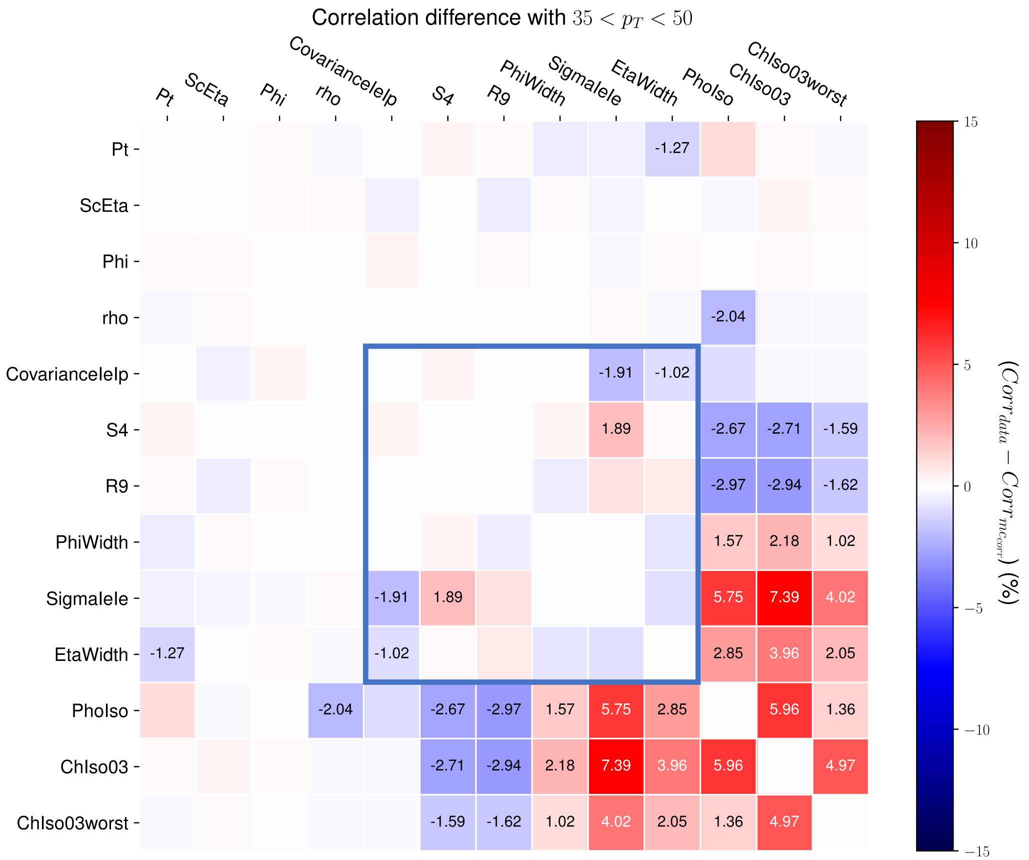


Closer agreement between data and MC corrected profiles

Correlations

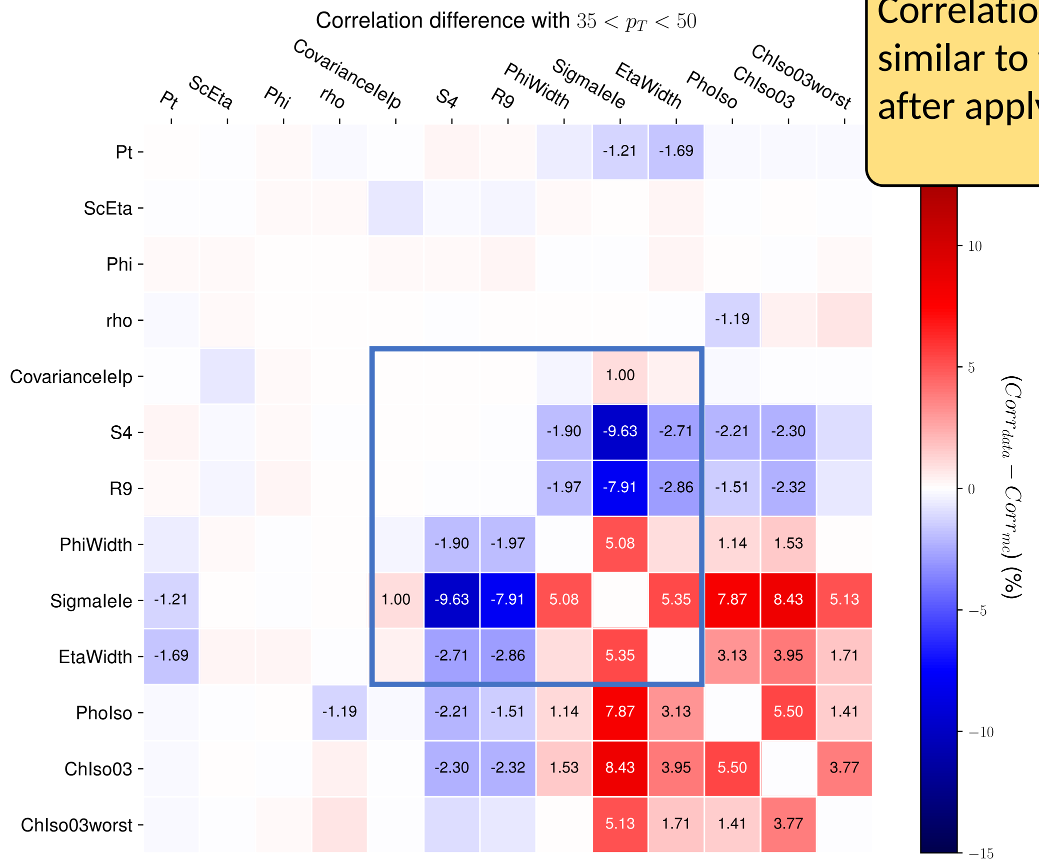
Difference of correlation matrices between data and MC **before** (left) and **after** (right) corrections:



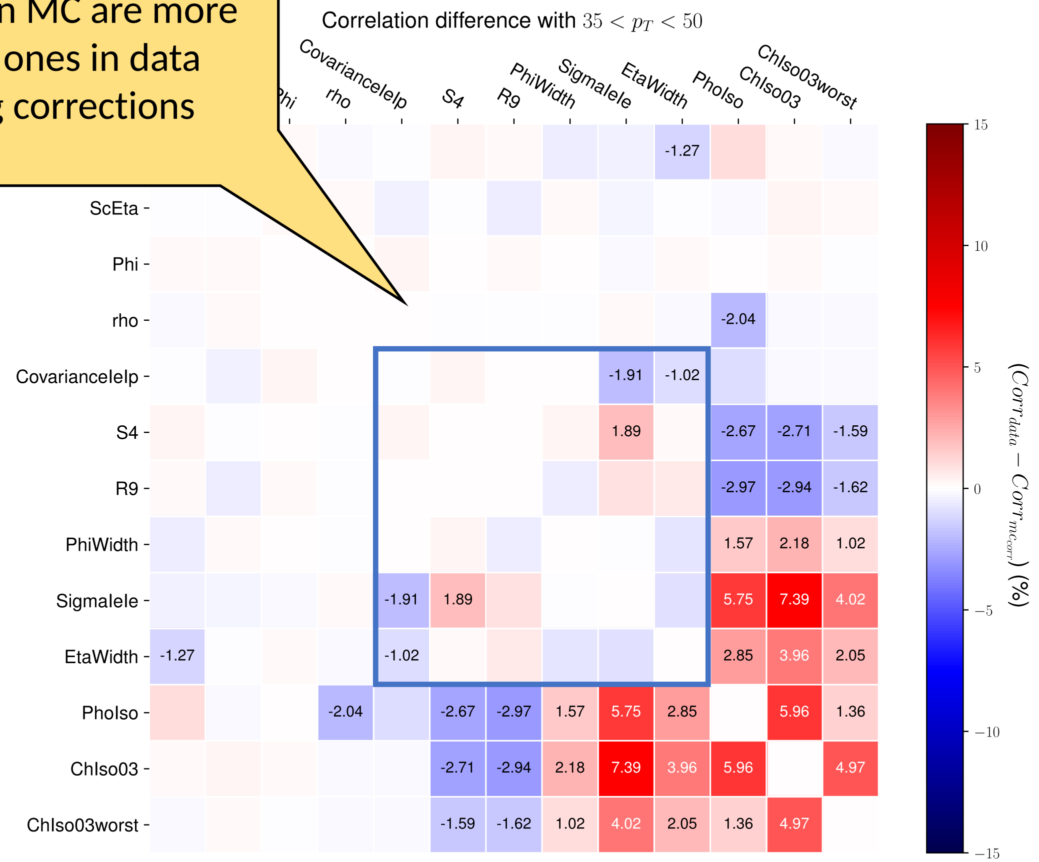


Correlations

Difference of correlation matrices between data and MC **before** (left) and **after** (right) corrections:

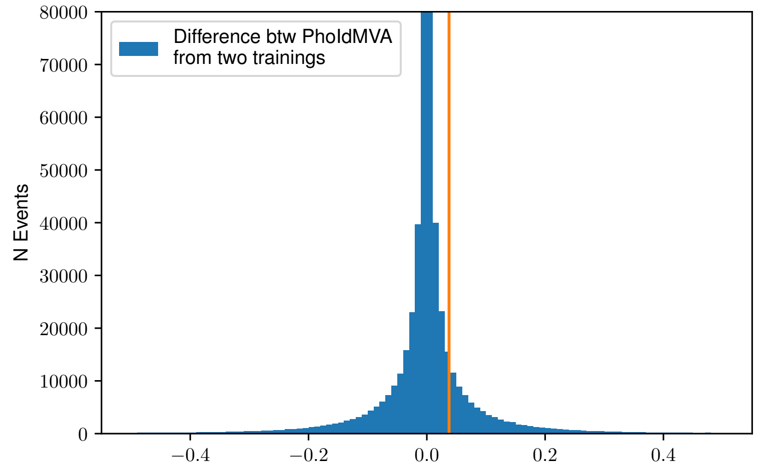


Correlations in MC are more similar to the ones in data after applying corrections

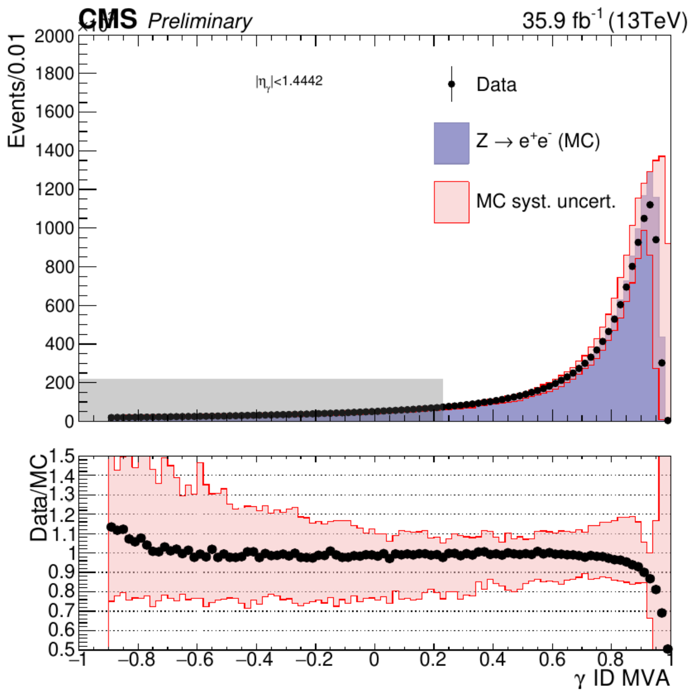


Systematic Uncertainties

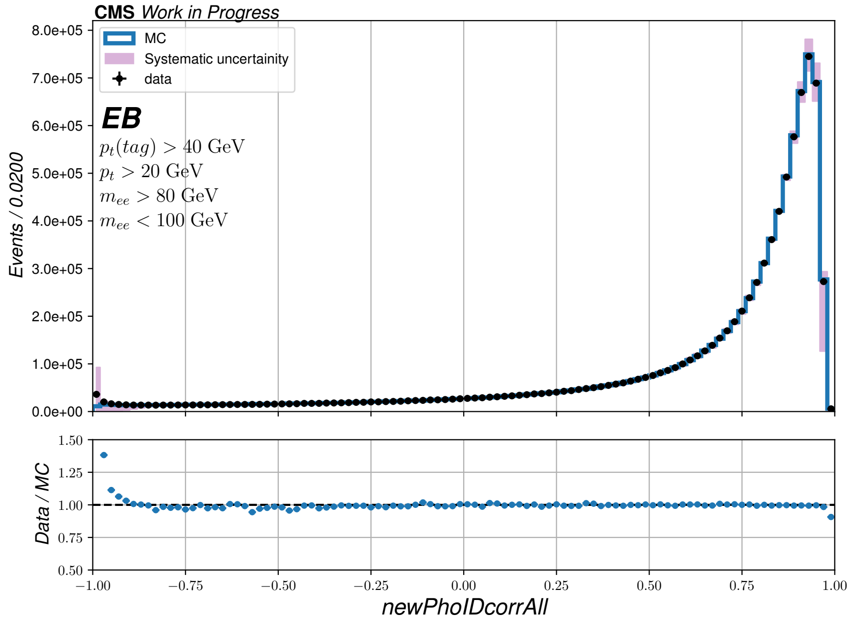
- Correction scheme accounts for all uncertainties and correlations → the only uncertainty comes from **finite size** of training sample
- Split the training sample in two and derive $\pm 1\sigma$ from the RMS of the $PhoID_1 - PhoID_2$ distribution



Run 1 Analysis (no corrections)



Run 2 Analysis



Future Prospects: Normalizing Flows

Also in its NN implementation, the CQR requires to train models and regressors one after the other to **take correlations into consideration**

Takes a long time and the corrected distributions need to be checked at every step

Normalizing flows allow to model high dimensional conditional distributions (see D. Valsecchi's talk)

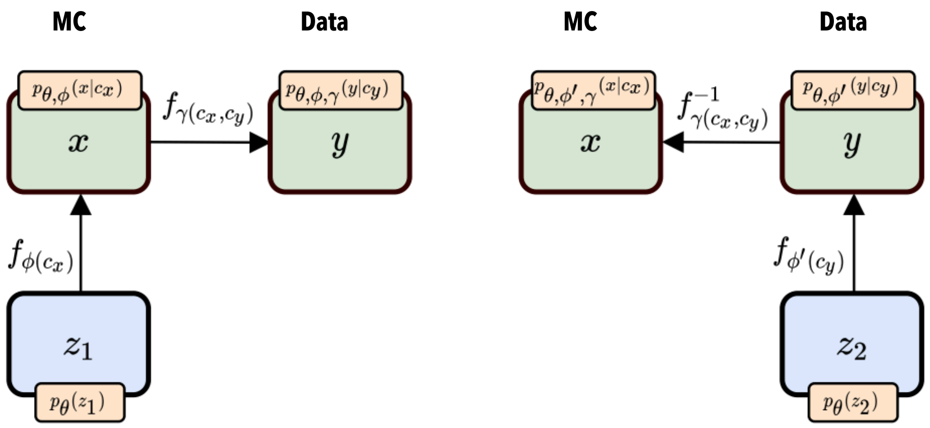
Benefit: chain is removed, making the training simpler and faster

CMS ETH zürich

Future Prospects: Normalizing Flows

Also in its NN implementation, the CQR requires to train models and regressors one after the other to **take correlations into consideration**

Takes a long time and the corrected distributions need to be checked at every step



From Flow4Flow paper

Normalizing flows allow to model high dimensional conditional distributions (see D. Valsecchi's talk)

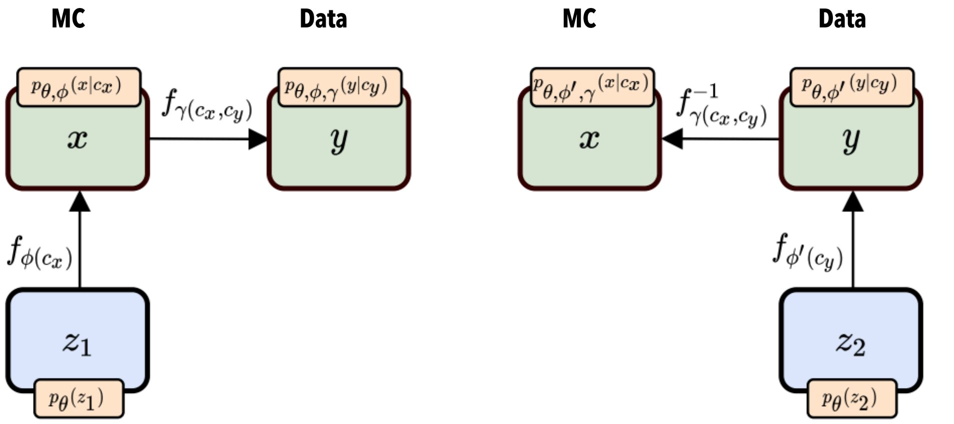
Benefit: chain is removed, making the training simpler and faster

As showed in arXiv:2211.02487 it is possible to train a system of three normalizing flows able to map two multidimensional conditional distributions into one another…
But does this procedure reach the level of precision that we require?

CMS ETH zürich

Future Prospects: Normalizing Flows

If you find this idea interesting and you want to work on it, we are offering this as a semester project - **Contact us!**



From Flow4Flow paper

As showed in arXiv:2211.02487 it is possible to train a system of three normalizing flows able to map two multidimensional conditional distributions into one another…
But does this procedure reach the level of precision that we require?

CMS ETH zürich

Thank you for your attention!

Backup

Correction Approach

- Developed procedure called **Chained Quantile Regression** (**CQR**) to match data with MC (and hence decrease systematic uncertainties)
- Corrections are derived using Tag & Probe method on $Z \rightarrow e^+e^-$ events, with the reconstruction of the probe leg as a photon
- PhotonID score is re-evaluated with corrected variables